Introduzione

A differenza di qualsiasi altro momento della storia, oggi hacker, informatori e archivisti sottraggono regolarmente terabyte di dati a governi, aziende e gruppi estremisti. Questi dataset spesso contengono vere "miniere d'oro" di rivelazioni di interesse pubblico, in molti casi rese liberamente disponibili al download da parte di chiunque. Eppure, questi "tomi" digitali possono rivelarsi estremamente difficili da analizzare o interpretare, e solo poche persone, oggi, hanno le competenze per farlo.

Ho scritto questo libro per i giornalisti, i ricercatori, gli hacktivisti e chiunque altro desideri apprendere le tecnologie e le competenze di programmazione necessarie per studiare questi dati hackerati o trapelati. Non presumo alcuna conoscenza preliminare da parte del lettore. Insieme alle lezioni sulla programmazione e sugli strumenti tecnici, ho incluso molti aneddoti e suggerimenti di prima mano dalle trincee del giornalismo investigativo. In una serie di progetti pratici, lavorerete con dataset reali, compresi quelli provenienti da dipartimenti di polizia, gruppi fascisti, milizie, una banda di ransomware russa e social network. Nel farlo, vi confronterete direttamente con il "campo di battaglia" rappresentato dagli eventi attuali del XXI secolo: l'ascesa del neofascismo e il rifiuto della realtà oggettiva, l'estrema divisione "partigiana" e un Internet traboccante di disinformazione.

Leggendo questo libro, acquisterete le competenze necessarie per scaricare e analizzare i vostri dataset, estraendo le rivelazioni in essi contenute e trasformando quelle informazioni quasi incomprensibili in veri report di primo piano.

Perché ho scritto questo libro

Lavoro dal 2013 come giornalista investigativo per *The Intercept*, riferendo su un'ampia varietà di dataset trapelati. Il primo dataset su cui mi sono fatto le ossa è stato lo *Snowden Archive*: una raccolta di documenti top-secret dell'informatore della NSA (*National Security Agency*) Edward Snowden, che rivelano che la NSA spia praticamente chiunque, nel mondo, usi un telefono o Internet. Ho scritto una decina di articoli e ho contribuito a pubblicare oltre duemila documenti segreti provenienti da quel dataset, contribuendo a portare le questioni della privacy e della sorveglianza governativa in primo piano nella coscienza pubblica e portando all'adozione diffusa di tecnologie di protezione della privacy.

Un tempo, queste enormi fughe di dati erano rare, ma oggi sono sempre più comuni. Nel mio lavoro presso *The Intercept*, mi imbatto in questi dataset così frequentemente che a volte mi sento come se stessi affogando nei dati; semplicemente ne ignoro la maggior parte, perché è impossibile, per me, esaminarli tutti. Sfortunatamente, spesso questo significa che nessuno ne parlerà e che i segreti che contengono rimarranno nascosti per sempre. Spero che questo libro aiuti a cambiare la situazione.

Le rivelazioni basate su dataset trapelati possono modificare il corso della storia. Nel 1971, la fuga di documenti militari da parte di Daniel Ellsberg, nota come *Pentagon Papers*, portò alla fine della guerra del Vietnam. Lo stesso anno, un gruppo di attivisti clandestini chiamato "Citizens' Commission to Investigate the FBI" ha fatto irruzione in una sede operativa del *Federal Bureau of Investigation*, ha sottratto documenti segreti e li ha fatti trapelare ai media. Questo dataset menziona COINTELPRO. Il giornalista della NBC Carl Stern ha utilizzato le richieste del *Freedom of Information Act* per rivelare pubblicamente che COINTELPRO era un'operazione segreta dell'FBI, dedita alla sorveglianza, all'infiltrazione e al discredito dei gruppi politici di sinistra. Questo dataset sottratto all'FBI ha portato anche alla creazione del "Church Committee", un comitato del Senato USA che ha indagato su questi abusi e li ha frenati. Più recentemente, le fughe di documenti di Chelsea Manning nel 2010 su Iraq e Afghanistan hanno contribuito a innescare la primavera araba, e documenti ed e-mail sottratti da hacker militari russi hanno contribuito all'elezione di Donald Trump come presidente degli Stati Uniti nel 2016.

A mano a mano che leggerete questo libro, eseguirete il download di una varietà di dataset reali, violati e trapelati, scoprendo come sono strutturati e come estrarne i segreti. Forse, un giorno, voi stessi cambierete la Storia. Leggerete anche storie da molti altri dataset, alcuni dei quali sono privati e non disponibili al download pubblico.

Che cosa scoprirete

Questo libro è diviso in cinque parti, ognuna delle quali si basa sulla parte precedente. Inizieremo con alcune considerazioni sulla sicurezza e sulla privacy, per esempio come verificare che i dataset siano autentici e come comunicare in modo sicuro con le fonti. Elaborerete poi i dataset nel terminale del vostro computer e su server remoti nel cloud e imparerete a rendere ricercabili i vari tipi di dataset e anche a cercare informazioni nei dump di e-mail. Seguirete un corso accelerato di programmazione Python, con particolare attenzione alla scrittura di codice per automatizzare le attività investigative. Queste competenze di programmazione vi consentiranno di analizzare dataset che contengono milioni di file, cosa impossibile da fare manualmente. Infine, tratterò due entusiasmanti casi di studio tratti dal mondo reale, in seguito ad alcune mie indagini.

Di seguito troverete in maggior dettaglio gli argomenti trattati, capitolo per capitolo.

Parte I – Fonti e dataset

La *Parte I* tratta i problemi che dovrete risolvere prima di iniziare ad analizzare i dataset: come proteggere le vostre fonti, come mantenere sicuri i vostri dataset e le vostre ricerche e come acquisire i dataset in modo sicuro.

Nel *Capitolo 1* imparerete a proteggere le vostre fonti dalle ritorsioni. Scoprirete come comunicare in modo sicuro con le fonti, come salvare i dataset riservati e come decidere

quali informazioni oscurare. Il capitolo tratta anche il passaggio fondamentale di come autenticare i dataset, utilizzando l'esempio dei registri delle chat di WikiLeaks e delle cartelle cliniche di un gruppo di estrema destra contrario alle vaccinazioni. Imparerete a proteggere la vostra vita digitale e, per estensione, a proteggere le indagini basate sui dati su cui state lavorando. Ciò include l'utilizzo di un gestore di password, la crittografia di dischi rigidi e dei dischi USB, la sanificazione di documenti potenzialmente dannosi con l'applicazione Dangerzone e altro ancora.

Nel *Capitolo 2* imparerete ad acquisire copie dei dataset violati e trapelati. Parlerò di DDoSecrets (*Distributed Denial of Secrets*), un collettivo per la trasparenza con cui sono coinvolto e che ospita copie di tutti i dataset con cui lavorerete in questo libro, e imparerete a scaricare i dataset da DDoSecrets utilizzando BitTorrent. Spiegherò diversi modi per acquisire dataset direttamente dalle fonti e introdurrò strumenti di sicurezza e anonimato come Signal, Tor Browser, OnionShare e SecureDrop. Per esempio, spiegherò come ho potuto comunicare con una fonte che ha fatto trapelare dati dal gruppo di attivisti conservatori Tea Party Patriots.

Eseguirete anche il download di una copia del dataset BlueLeaks, uno dei principali dataset con cui lavorerete in questo libro. BlueLeaks è una raccolta di 270 GB di dati trapelati da centinaia di siti web delle forze dell'ordine statunitensi nell'estate del 2020, nel bel mezzo della rivolta di Black Lives Matter. Come vedrete, è pieno di prove di cattiva condotta della polizia. BlueLeaks è stato ampiamente trattato dalla stampa, ma la maggior parte dei dati non è ancora emersa. Alla fine di questo libro avrete tutti gli strumenti necessari per condurre le vostre indagini su BlueLeaks.

Parte II – Gli strumenti del mestiere

Nella *Parte II* vi eserciterete nell'uso dell'interfaccia a riga di comando per valutare rapidamente i dataset trapelati e per utilizzare strumenti che non dispongono di un'interfaccia grafica, sviluppando competenze che applicherete ampiamente nel resto del libro.

Nel *Capitolo 3* apprenderete le nozioni di base per controllare il vostro computer tramite i comandi testuali, alla riga di comando, oltre a vari suggerimenti e trucchi per valutare ed eseguire ricerche rapidamente nei dataset come BlueLeaks utilizzando solo la riga di comando. Scriverete anche il vostro primo script della shell, un file contenente una sequenza di comandi testuali.

Nel *Capitolo 4* amplierete le vostre competenze di base sulla riga di comando, apprendendo nuovi comandi e configurando nel cloud un server per analizzare in remoto i dataset violati e trapelati. Per esempio, lavorerete con il dataset Oath Keepers, che contiene e-mail dei miliziani di destra che hanno partecipato a una cospirazione sediziosa con lo scopo di mantenere Trump al potere dopo la sconfitta elettorale del 2020.

Nel *Capitolo 5* imparerete a utilizzare Docker, una tecnologia che consente di utilizzare un'ampia varietà di software complessi, fondamentali per l'analisi dei dataset. Utilizzerete Docker per eseguire Aleph, un software in grado di analizzare dataset di grandi dimensioni, trovare legami e cercare parole chiave nei dati.

Il *Capitolo 6* si concentra sugli strumenti e sulle tecniche per eseguire indagini sui dump di e-mail. Leggerete le e-mail delle forze di polizia di Nauru sui centri di detenzione offshore dell'Australia, inclusi molti messaggi sui rifugiati che cercano asilo in Australia, e dello stesso presidente di Nauru. Indagherete anche sulle e-mail di un think tank conservatore chiamato Heritage Foundation, che include argomenti omofobici contro il

matrimonio gay. Utilizzando le competenze apprese, sarete in grado di eseguire ricerche sui dump di posta elettronica che potrete acquisire in futuro.

Parte III – Programmazione Python

Nella *Parte III* seguirete un corso accelerato di programmazione in linguaggio Python, incentrato sulle competenze necessarie per analizzare i dataset violati e trapelati trattati nei capitoli successivi.
Il *Capitolo 7* introduce i concetti di base della programmazione: imparerete a scrivere ed eseguire script e comandi Python nell'interprete interattivo di Python, a fare calcoli, a definire variabili, a utilizzare le stringhe e la logica booleana, a scorrere liste di elementi e a utilizzare funzioni.
Il *Capitolo 8* si basa sui fondamenti di Python trattati in precedenza. Imparerete a esplorare i filesystem e a lavorare con dizionari e liste. Infine, metterete in pratica la teoria scrivendo diversi script Python per aiutarvi a indagare su BlueLeaks ed esplorare i registri delle chat trapelate dal gruppo di ransomware russo Conti.

Parte IV – Dati strutturati

Nella *Parte IV* imparerete a lavorare con alcuni dei formati di file più comuni nei dataset violati e trapelati.
Nel *Capitolo 9* conoscerete la struttura del formato CSV (valori separati da una virgola), visualizzando i file CSV all'interno sia di software per fogli di lavoro sia di editor di testo. Preparerete quindi script Python per scorrere le righe di un file CSV e per salvare i vostri file CSV, consentendovi di analizzare ulteriormente i file CSV nel dataset BlueLeaks.
Il *Capitolo 10* introduce un'applicazione personalizzata, *BlueLeaks Explorer*, che ho sviluppato e rilasciato insieme a questo libro: vi descrivo come e perché ho creato l'app e vi mostro come utilizzarla. Potete utilizzare questa app per indagare sulle numerose parti del dataset BlueLeaks che non sono ancora state analizzate, alla ricerca di nuove rivelazioni sulle agenzie di intelligence negli Stati Uniti. Se mai aveste bisogno di sviluppare un'app per analizzare un determinato dataset, potrete utilizzare come fonte di ispirazione questo capitolo.
Il *Capitolo 11* si concentra sul formato JSON e sul dataset Parler, contenente oltre un milione di video caricati sul sito di social network di estrema destra Parler, inclusi migliaia di video dell'insurrezione del 6 gennaio 2021 al Campidoglio degli Stati Uniti. Questo dataset include metadati in formato JSON per ogni video, comprese informazioni relative a quando è stato girato e in quale luogo. Alcuni di questi video sono stati utilizzati come prova durante la seconda inchiesta di impeachment per Donald Trump. Preparerete script Python per filtrare questi video e tracciare le coordinate GPS dei video di Parler su una mappa, in modo da poter utilizzare dati di geolocalizzazione simili nelle successive indagini.
Nel *Capitolo 12* imparerete a estrarre rivelazioni dai database SQL lavorando con il dataset Epik. Epik è una società di orientamento nazionalista cristiano che fornisce nomi di dominio e servizi di web hosting a esponenti di estrema destra, compresi siti noti per ospitare i manifesti degli attentatori di massa. Il dataset Epik contiene enormi database pieni di dati sui clienti violati, insieme a informazioni sulla vera proprietà dei nomi di

dominio per i siti web estremisti, informazioni che dovrebbero essere nascoste dietro un servizio di privacy dei nomi di dominio. Impiegherete le vostre nuove competenze per scoprire nomi di dominio di proprietà di una delle persone che si celano dietro QAnon e il forum di immagini di estrema destra 8kun. Se siete interessati alle indagini sull'estremismo, il dataset Epik potrebbe essere utile per le vostre indagini future.

Parte V – Casi di studio

La *Parte V* tratta due casi di studio che ho approfondito nella mia carriera, descrivendo come ho condotto importanti indagini utilizzando le competenze apprese. In entrambi, spiego il mio iter investigativo: come ho ottenuto i miei dataset, come li ho analizzati utilizzando le tecniche descritte nel libro, quale codice Python ho scritto per aiutarmi in questa analisi, quali rivelazioni ho scoperto e quale impatto sociale ha avuto la mia attività giornalistica.

Nel *Capitolo 13* discuto la mia indagine sugli AFLDS (*America's Frontline Doctors*), un gruppo di destra contrario alle vaccinazioni nato durante la pandemia di Covid-19 per opporsi alle misure di sanità pubblica. Spiegherò come ho trasformato una raccolta di file CSV e JSON violati in un importante report, rivelando che una rete di losche società di telemedicina ha truffato decine di milioni di dollari a coloro che erano scettici sui vaccini. Il mio report ha portato a un'indagine del Congresso su AFLDS.

Nel *Capitolo 14* descrivo come ho analizzato e pubblicato enormi quantità di dati di registri trapelati di chat di neonazisti. Parlo anche del mio ruolo nello sviluppo di uno strumento di indagine pubblica per tali dataset: *DiscordLeaks*. Questo strumento ha contribuito a una causa di successo contro gli organizzatori del sanguinoso raduno *Unite the Right* nel 2017, che ha portato a un risarcimento di oltre 25 milioni di dollari contro i leader del movimento fascista americano.

Appendici

L'*Appendice A* include suggerimenti per gli utenti Windows che svolgono gli esercizi di questo libro, per facilitare l'esecuzione del codice. L'*Appendice B* parla del *web scraping*: come scrivere codice che accede ai siti web per voi in modo da automatizzare il vostro lavoro investigativo o creare i vostri dataset da siti web pubblici.

Di che cosa avrete bisogno

Questo libro è un tutorial interattivo: ogni capitolo include esercizi, tranne i casi di studio della Parte V. Molti esercizi richiedono che abbiate completato gli esercizi precedenti, quindi vi consiglio di leggere questo libro in modo sequenziale. Per esempio, nel Capitolo 1, crittograferete un disco USB sul quale, nel Capitolo 2, eseguirete il download di una copia del dataset BlueLeaks.

Leggete questo libro con il computer acceso, completando gli esercizi e provando le tecnologie e il software a mano a mano che li imparate. Il codice sorgente di ogni esercizio, così come il codice utilizzato nei casi di studio e nelle appendici, è disponibile al

download in un repository online organizzato a capitoli su `https://github.com/micahflee/hacks-leaks-and-revelations` e sul sito di Apogeo all'indirizzo `https://bit.ly/apo-hfdr`.
Per rendere questo libro il più possibile accessibile a tutti, ho cercato di mantenere i requisiti semplici e limitati. Avrete bisogno di quanto segue.

- *Un computer Windows, macOS o Linux*. Windows è molto diverso da macOS e Linux, ma spiegherò tutti i passaggi aggiuntivi con i quali gli utenti Windows potranno configurare il proprio computer. Se siete utenti Linux, presumo che stiate utilizzando Ubuntu; se state utilizzando un'altra distribuzione di Linux, potreste dover modificare leggermente i comandi.
- *Un disco USB con almeno 1 TB di capacità*. Lo utilizzerete per salvare i dataset di grandi dimensioni con cui lavorerete.
- *Una connessione Internet in grado di scaricare circa 280 GB di dataset e molti altri gigabyte di software.* Se vivete in un paese dotato di un servizio Internet minimamente "decente", una comune connessione Internet domestica dovrebbe andare bene, anche se potrebbero essere necessarie ore o giorni per scaricare i dataset più grandi nel libro. In alternativa, potreste trovare connessioni Internet più veloci nelle biblioteche locali, nei bar o nei campus universitari.
- Per i due esercizi in cui lavorerete con dataset provenienti da server nel cloud, vi serviranno anche *alcuni dollari o euro e una carta di credito* per pagare il provider di cloud-hosting.

Ora prendete il vostro laptop, il vostro disco USB e magari un caffè o un tè e preparatevi a iniziare la vostra caccia a nuove rivelazioni.

L'autore

Micah Lee è un giornalista investigativo ed esperto di sicurezza informatica, noto per aver protetto la diffusione delle sconcertanti rivelazioni di Edward Snowden sulla NSA. È direttore della sicurezza informatica della rivista *The Intercept* e consulente del collettivo di giornalisti *Distributed Denial of Secrets* dedicato alla divulgazione di notizie di pubblico interesse. È inoltre cofondatore della Freedom of the Press Foundation, collaboratore del Tor Project e sviluppatore di strumenti di sicurezza open source come OnionShare e Dangerzone.

La revisione tecnica

Jennifer Helsby è core engineer presso Penumbra Labs, dove si occupa di crittografia applicata. In precedenza, è stata direttrice dello sviluppo del progetto SecureDrop, che rende possibili comunicazioni private e anonime tra i giornalisti e le loro fonti, e ricercatrice post-doc presso il Center for Data Science and Public Policy dell'Università di Chicago. Ha conseguito un dottorato in astrofisica presso l'Università di Chicago nel 2015.